DEPARTMENT OF MATHEMATICS AND STATISTICS
UMASS - AMHERST
BASIC EXAM - STATISTICS
August 2013

Work all problems. Show all work. Explain your answers. State the theorems used whenever possible. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level.

1. Suppose independent identically distributed random variables Z_i with distribution given by:

$$Y_i \sim Poisson(\lambda)$$
$$Z_i \sim Neg.Binom(Y_i, p),$$

where the Y_i's are also independent of each other, the Y_i's follow the standard Poisson distribution, and the negative binomial distribution is parameterized such that

$$P(Z_i = z_i|Y_i) = \binom{z_i + Y_i - 1}{z_i}(1-p)^{Y_i}p^{z_i}$$

and

$$E(Y_i) = V(Y_i) = \lambda, \quad E(Z_i|Y_i) = \frac{pY_i}{1-p}, \quad V(Z_i|Y_i) = \frac{pY_i}{(1-p)^2}.$$

   (a) (6 points) Find $E(Z_i)$ and $V(Z_i)$.

   (b) (6 points) Given observations $Z_1, Z_2, \ldots, Z_n$, give expressions for the method of moments (MOM) estimators for p and λ.

   (c) (6 points) Consider possible datasets with the same sample mean, but different sample variance. Draw a sketch of the MOM estimates for p and λ as a function of the sample variance. You do not need to include a numerical scale on the axes, but do illustrate the general trends.

   (d) (6 points) Write down the joint distribution of $Z_1, Z_2, \ldots, Z_n$ and $Y_1, Y_2, \ldots, Y_n$ (denoted as $f(Z_1, Z_2, \ldots, Z_n, Y_1, Y_2, \ldots, Y_n|p, \lambda)$), and the marginal distribution of $Z_1, Z_2, \ldots, Z_n$, the latter of which can be left as a summation.

   (e) To obtain the MLEs of p and λ, the marginal distribution of the observations $Z_1, Z_2, \ldots, Z_n$ should be maximized. As the marginal distribution does not have an explicit form, the EM (expectation-maximization) algorithm can be used to obtain the MLEs.

      (a) (5 points) In the E step, the following function is obtained

$$Q(p, \lambda|p_{old}, \lambda_{old}) = E(\log(f(Z_1, Z_2, \ldots, Z_n, Y_1, Y_2, \ldots, Y_n|p, \lambda))|Z_1, Z_2, \ldots, Z_n, p_{old}, \lambda_{old}).$$

      Show that the Q function here, as a function of p and λ is proportional to

$$Q(p, \lambda|p_{old}, \lambda_{old}) \propto \log(1-p)\sum_{i=1}^{n} E(Y_i|Z_i, p_{old}, \lambda_{old}) + \log p \sum_{i=1}^{n} Z_i - n\lambda + \log\lambda \sum_{i=1}^{n} E(Y_i|Z_i, p_{old}, \lambda_{old}),$$

      where $E(Y_i|Z_i, p_{old}, \lambda_{old})$ is the conditional expection of Y_i given Z_i evaluated at $p = p_{old}$ and $\lambda = \lambda_{old}$, some initial or previous values of p and λ.

      (b) (5 points) In the M step, the Q function is maximized in terms of p and λ to obtain their updated values. Show that the updated values, denoted as p_{new} and λ_{new}, are

$$p_{new} = \frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} E(Y_i|Z_i, p_{old}, \lambda_{old}) + \sum_{i=1}^{n} Z_i}, \quad \lambda_{new} = \frac{\sum_{i=1}^{n} E(Y_i|Z_i, p_{old}, \lambda_{old})}{n}.$$

2. Let $X_1, \cdots, X_n$ be independently and identically distributed as $N(\theta, 1)$. Define

$$Y_i = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{if } X_i \leq 0 \end{cases} .$$

Let $\psi = P(Y_1 = 1)$.

(a) (6 points) Find the maximum likelihood estimator $\hat{\psi}$ of ψ based on $X_1, \cdots, X_n$. *Hint: Write your solution in terms of the cumulative function $\Phi(x)$ of the standard normal distribution.*

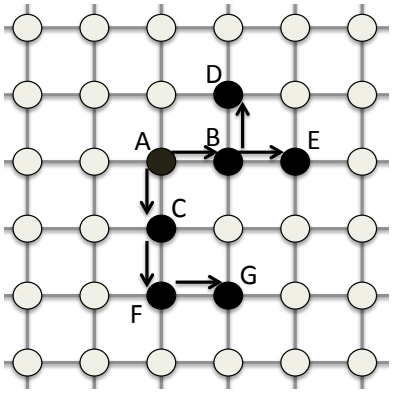
(b) (6 points) Find an approximate 95 percent confidence interval for ψ based on $X_1, \cdots, X_n$.

(c) (6 points) Define $\tilde{\psi} = (1/n)\sum_i Y_i$. Show that $\tilde{\psi}$ is a consistent estimator of ψ.

(d) (6 points) Compute the asymptotic relative efficiency of $\tilde{\psi}$ to $\hat{\psi}$. Express your answer in terms of θ. *Hint: Use the delta method to get the standard error of the MLE. Then compute the standard error of $\tilde{\psi}$.*

3. A disease spreads by contact along a regular grid network, part of which is shown below. Each link transmits infection independently with probability p.

Suppose A became infected, and the infection spread according to the arrows in the diagram.



(a) (6 points) Find an expression for the probability of this spread in terms of p, given A was infected first.

(b) (6 points) Express the answer to the previous part as a likelihood for p, and find the MLE for p.

(c) (6 points) Consider an arbitrary observed spread pattern on the grid above. Find the minimal sufficient statistic for p. You may use T to represent the number of transmission events and $\bar{T}$ to represent the number of possible infections that did not occur. Is the statistic in the previous part complete? Why or why not?

4. Let $X_1, \cdots, X_n$ be independently and identically distributed as $N(\theta, 1)$. Consider testing

$$H_0 : \theta = 0 \text{ versus } \theta \neq 0.$$

(a) (6 points) Consider the test that has the rejection region $R = \{x^n = (x_1, \cdots, x_n) : |\bar{x}^n| > c\}$. Here $\bar{x}^n = (1/n)\sum_i x_i$. Find c so that the test has size α.

(b) (6 points) Find the power of the test in (a) at $\theta = 1$.

(c) (6 points) Find the p-value of the test in (a) when the observed data are $x_1, \cdots, x_n$.

(d) (6 points) Now consider the Bayesian test that rejects H_0 if $P(H_0|X_1 = x_1, \cdots, X_n = x_n) < P(H_1|X_1 = x_1, \cdots, X_n = x_n)$. Let the prior for the hypothesis be $P(H_0) = P(H_1) = 1/2$. Let the prior for θ under H_1 be $\theta \sim N(0, b^2)$. Find an expression for $P(H_0|X_1 = x_1, \cdots, X_n = x_n)$. No need to show this: one can see that the posterior probability of H_0 can be large even when the p-value is small, especially when n is large. This disagreement between Bayesian and frequentist testing is called the Jeffreys−Lindley paradox.